Journal of Southeast University (English Edition) Vol. 35, No. 3, pp. 393 - 398 Sept. 2019 ISSN 1003—7985

Travel time prediction model of freeway based on gradient boosting decision tree

Cheng Juan Chen Xianhua

(School of Transportation, Southeast University, Nanjing 211189, China)

Abstract: To investigate the travel time prediction method of the freeway, a model based on the gradient boosting decision tree (GBDT) is proposed. Eleven variables (namely, travel time in current period T_i, traffic flow in current period Q_i, speed in current period V_i, density in current period K_i, the number of vehicles in current period N_i, occupancy in current period R_i, traffic state parameter in current period X_i, travel time in previous time period T_{i-1}, etc.) are selected to predict the travel time for 10 min ahead in the proposed model. Data obtained from VISSIM simulation is used to train and test the model. The results demonstrate that the prediction error of the GBDT model is smaller than those of the back propagation (BP) neural network model and the support vector machine (SVM) model. Travel time in current period T_i is the most important variable among all variables in the GBDT model. The GBDT model can produce more accurate prediction results and mine the hidden nonlinear relationships deeply between variables and the predicted travel time.

Key words: gradient boosting decision tree (GBDT); travel time prediction; freeway; traffic state parameter

DOI: 10.3969/j.issn.1003-7985.2019.03.017

Travel time is the most intuitionistic index to reflect the running condition, which is an important foundation for constructing intelligent transportation systems (ITS)[1]. With accurate travel time information, on the one hand, travelers can make better travel choices; on the other hand, traffic managers can improve traffic management decisions[2].

There are many methods for predicting the travel time, such as mathematical statistics methods[3-4] and machine learning methods[5]. Since the travel time prediction has typical nonlinear characteristics, the travel time prediction based on machine learning methods is more accurate than the methods based on mathematical statistics. Therefore, the travel time prediction method has gradually transferred to machine learning methods, such as artificial neural networks[6-7], support vector machines (SVM)[8-10], Kalman filters[11], the kernel-clustering algorithm[12] and the K nearest method[13-14]. A variety of models are proposed based on the methods. However, it is difficult for traffic researchers and managers to explain the relationship between indicators and the predicted travel time through these models. In view of this, the gradient boosting decision tree (GBDT) is used to build the travel time prediction model in this paper. GBDT combines the advantages of data mining to dig deep into the impact of variables on the predicted travel time.

The GBDT model provides a flexible framework to adopt different types of predictors as the input variables (for instance, traffic flow, speed, density, occupancy, number of vehicles, traffic state parameter and data-time variables). Meanwhile, the GBDT model understands the diverse influences of different variables on the predicted travel time, explores the nonlinear relationship between variables and the predicted travel time, and has good interpretability.

1 Travel Time Prediction Model Based on GBDT

GBDT is an iterative decision tree algorithm, which is based on the idea of boosting iteration. The foundation of GBDT is the classification and regression tree (CART) algorithm. Except for the first decision tree generated using the original indicator, the target in each iteration minimizes the loss function value of the current learner, that is, the loss function always falls along its gradient direction. By successive iterations, the final residual approaches zero. The results of all trees are added up as the final prediction result[15-17].

Suppose that $X^i = \{x_i^1, x_i^2, \ldots, x_i^k\}$ is the K-dimensional variable that affects travel time. y^i is the response variable of the travel time, namely the target variable. For N training samples $\{(X^1, y^1), (X^2, y^2), \ldots, (X^N, y^N)\}$, the GBDT modeling process is described as follows.

1) Initialize the learner, that is

$$f_0(X) = \arg\min_c \sum_{i=1}^{N} L(y_i, c) \tag{1}$$

where $f_0(X)$ is the initial decision tree with only one root node; $L(y_i, c)$ is the loss function; y_i is the i-th training data; c is a constant value that minimizes the loss function.

Received 2018-08-30, **Revised** 2019-05-28.

Biographies: Cheng Juan (1983—), female, Ph. D. candidate; Chen Xianhua (corresponding author), male, doctor, professor, chenxh@seu.edu.cn.

Foundation item: The National Natural Science Foundation of China (No. 51478114, 51778136).

Citation: Cheng Juan, Chen Xianhua. Travel time prediction model of freeway based on gradient boosting decision tree[J]. Journal of Southeast University (English Edition), 2019, 35(3): 393 - 398. DOI: 10.3969/j.issn.1003-7985.2019.03.017.

Travel time prediction is a regression problem. In GBDT, there are many loss functions for the regression problem, such as the squared loss function, absolute value loss function, and Huber loss function. The squared loss function used in the GBDT model is

$$L(y, f(x)) = \frac{1}{2}[y - f(x)]^2 \tag{2}$$

where $f(x)$ is the learner obtained from the current iteration.

2) Let the number of iterations be $m = 1, 2, \ldots, M$, and the negative gradient of the i-th training data is

$$g_{mi} = -\left[\frac{\partial L(y, f(x_i))}{\partial f(x_i)}\right]_{f(x) = f_{m-1}(x)} \tag{3}$$

According to all samples and their negative gradient direction $(x_i, g_{mi})(i = 1, 2, \ldots, N)$, a decision tree T_m consisting of J leaf nodes is obtained. The j-th leaf node region is $R_{mj}(j = 1, 2, \ldots, J)$. The best residual fitting value of each leaf is

$$c_{mj} = \arg\min_{c} \sum_{x \in R_{mj}} L(y_i, f_{m-1}(x_i) + c) \tag{4}$$

The learner obtained in this iteration is

$$f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J} c_{mj} I \qquad x \in R_{mj} \tag{5}$$

where $I(x_i \in R_{mj})$ is the explanatory function of the i-th training data in the j-th leaf node region, and

$$I = \begin{cases} 1 & x_i \in R_{mj} \\ 0 & x_i \notin R_{mj} \end{cases}$$

3) After the m-th iteration, the final model is expressed as

$$f(x) = f_M(x) = c + \sum_{m=1}^{M} \sum_{j=1}^{J} c_{mj} I \qquad x \in R_{mj} \tag{6}$$

Via the times that a variable appears in the decision tree and the performance of the model after each segmentation, the variable importance of the model can be obtained[17].

$$I_k^2(F) = \frac{1}{M} \sum_{m=1}^{M} I_k^2(T_m) \tag{7}$$

$$I_k^2(T_m) = \sum_{j=1}^{J} E_j^2 1_j(X^k) \tag{8}$$

where T_m is the m-th decision tree in GBDT F with J leaf nodes; $1_j(X^k)$ is the indicator function that the variable X^k is chosen as split variable at node j in decision tree T_m; E_j^2 is the squared error improvement of the corresponding node after selecting variable X^k to split; $I_k^2(F)$ is the importance value of variable X^k in GBDT F; $I_k^2(T_m)$ is the importance value of variable X^k in the decision tree T_m.

2 Data

In this paper, the VISSIM simulation software developed by PTV is used to analyze the travel time in the freeway. The length of 1 048.28 m between the airport interchange and Lukou interchange is selected as the research area. Time detectors are set at both ends of the selected freeway section. The route diagram is presented in Fig. 1.

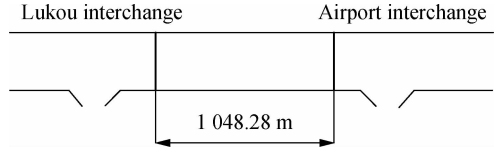

Fig. 1 The study area

VISSIM simulation software is calibrated according to actual hourly traffic flow on the Nanjing Airport freeway from Nanjing to the Airport investigated at a airport toll station from 9:00 to 15:00 on August 22, 2017. Since the actual traffic flow does not include congestion, in order to cover the states of free-flow, transition and congestion in the freeway, the traffic flow is increased by 600 veh/h from the actual measured value of the previous period during 15:00—17:00, which reflects the congestion state. Only increasing the number of vehicles cannot lead to congestion. However, based on the state of transition, the authors guarantee that all variables are constant, and continue to increase the traffic flow to characterize the state of congestion.

Through investigation, the vehicle proportion of car, truck, bus and taxi on the freeway is 0.42∶0.12∶0.26∶0.2.

In the freeway, the expected speed distributions of car, truck, and bus is 120, 100, and 100 km/h. The speed distributions of cars, trucks, buses, and taxis are shown in Fig. 2.

Using different random seed numbers, the experiment is simulated 133 times and the simulation time is 28 800 s. Finally, 133 sets of data are obtained, which represent 133 days' data of 9:00—17:00. Data of 133 d are divided into two data sets, in which 27 to 133 d of data are used as training data sets and 1 to 26 d of data are used as test data sets.

The travel time is obtained at the sampling interval of 300 s. T_i is used to represent the travel time at time step i (i is the current period), where $i = 1, 2, \ldots, 93$, represents 93 time periods from 9:15 to 17:00 (The first three periods 9:00 to 9:15 are taken as the pretreatment time of VISSIM). Considering the short-term prediction of travel time, the prediction period is set to be 10 min ahead.

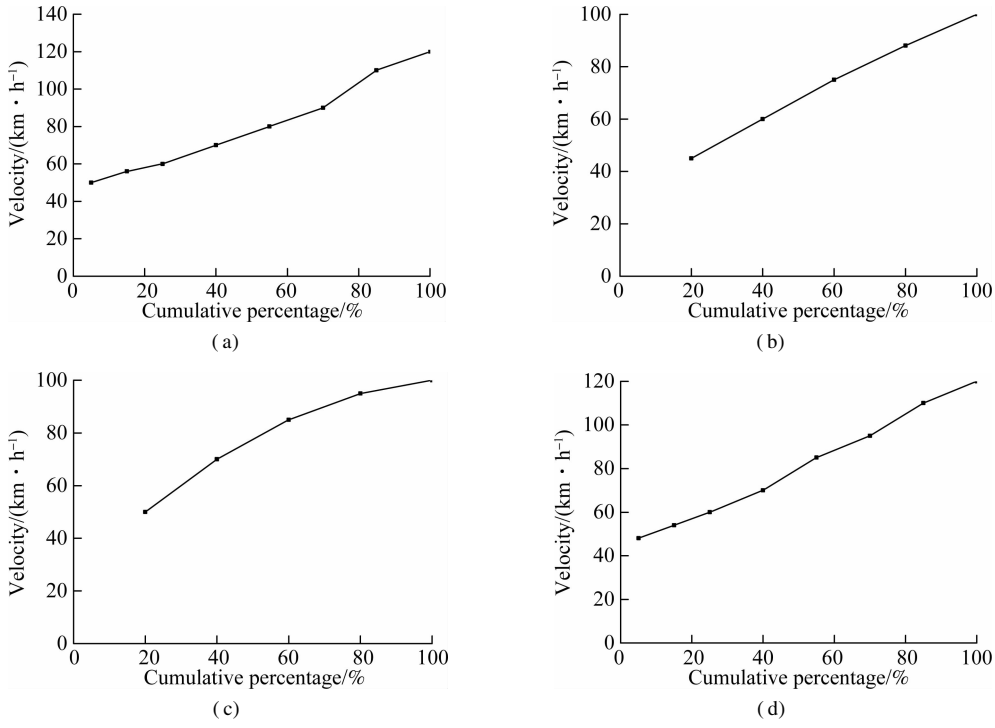

Fig. 2 Speed distribution. (a) Cars; (b) Trucks; (c) Buses; (d) Taxis

3 Establishment and Verification of the Model

3.1 Variables of the model

3.1.1 Traffic state parameter

In the Highway Capacity Manual[18], the freeway traffic state is divided into six grades (namely A to F) by means of average speed and density. As we all know, speed, density, and traffic flow are three basic parameters, which are interrelated. If the values of two parameters are known, the third one can be calculated. The standard of traffic state classification of a freeway is shown in Tab. 1.

Tab. 1 The standard of traffic state classification of a freeway[18]

Traffic state	Density range/ (vehicle · (km · lane)$^{-1}$)	Design speed/(km · h^{-1})			
		120		100	
		Speed/ (km · h^{-1})	Traffic flow/ (vehicle · (h · lane)$^{-1}$)	Speed/ (km · h^{-1})	Traffic flow/ (vehicle · (h · lane)$^{-1}$)
A	17.7	120.7	820	104.6	710
B	29.0	120.3	1 350	104.6	1 170
C	41.8	113.6	1 830	103.9	1 680
D	56.3	100.1	2 170	96.1	2 090
E	72.4	85.8	2 400	83.9	2 350
F	>72.4	<85.8	>2 400	<83.9	>2 350

In this study, traffic state parameters refer to the standard of traffic state classification of the freeway, let $x = 1, 2, \cdots, 6$ represent traffic states A to F, respectively. This paper combines existing traffic state levels and describes the freeway at a low level. Therefore, the traffic state of the freeway is divided into three categories. The free-flow state includes traffic states A and B, namely, $x_f = 1, 2$. The transition state includes traffic states C and D, namely $x_t = 3, 4$. The congestion state includes traffic states E and F, namely $x_c = 5, 6$. The traffic state parameter is $X = \{x_f, x_t, x_c\}$.

The travel time is affected by traffic states. In order to clarify the impact of traffic states on travel time prediction, the traffic state parameter in current period X_i is introduced into the GBDT model.

3.1.2 Variables of the model

Traffic flow, speed, and density are three basic parameters that characterize traffic flow characteristics and affect the travel time of the vehicle. In addition, occupancy and the number of vehicles also have a certain impact on travel time. Therefore, traffic flow in current period Q_i, speed in current period V_i, density in current period K_i, occupancy in current period R_i and the number of vehicles in current period N_i are introduced as input variables.

Other factors that have been discussed in previous studies[19] are also considered, that is, T_i is the travel time in

current period; T_{i-1} is the travel time at time step $i-1$; T_{i-2} is the travel time at time step $i-2$; ΔT_i is the change of travel time over two adjacent time steps, $\Delta T_i = T_i - T_{i-1}$; ΔT_{i-1} is the changes of travel time over two adjacent time steps, $\Delta T_{i-1} = T_{i-1} - T_{i-2}$.

The target variable of the model, namely the predicted travel time, is the travel time at time step $i+1$, which is denoted as T_{i+1}.

3.2 Results and verification of the model

In GBDT, there are five parameters that need to be determined, namely, the number of leaf nodes in a single regression tree J, learn rate η, the amount of attribute sampling S_a, subsample fraction f, and the number of regression trees M. The paper uses data mining software Salford systems developed by the Salford Company of the United States to establish the GBDT model. After repeated experiments, all the parameters of the GBDT model are obtained, that is $\{J, \eta, S_a, f\} = \{9, 0.01, 9, 0.6\}$. The optimal number of regression trees based on the minimum value of the objective function is automatically determined. The number of regression trees is 923.

The training data sets are used to train the model, and the test data sets are used for testing. The results show that the error of the model in the training data sets is 3.59%, and that in the test data sets is 3.94%.

To test the effectiveness of the GBDT model, the back propagation (BP) neural network model with three-layer feedforward perceptron algorithm and the SVM model with radial basis function (RBF) as the kernel function are also established by using the same training data sets. Then, the same test data sets are used for testing. Tab. 2 is the error of different models.

Tab. 2 MAPE of different models %

Data set	GBDT	BP neural network	SVM
Training data	3.59	4.49	6.47
Test data	3.94	4.70	6.54

3.3 Analysis of variables

The importance values of variables are determined in the GBDT model by the times of variables appearing in the decision tree. The relative importance value of the GBDT model is indicated in Fig. 3. It can be seen from Fig. 3 that the most important influence variable is the travel time in current period T_i. The travel time of the current period has the greatest influence on the travel time of the next period. As expected, the immediate previous traffic state will influence traffic in the near future. The influences of R_i, ΔT_i, N_i, and ΔT_{i-1} on the model are relatively small, indicating that the occupancy and the number of vehicles cannot directly affect the predicted travel time. The influence of the time difference on the model is less than that of the travel time of the two periods close to the predicted travel time.

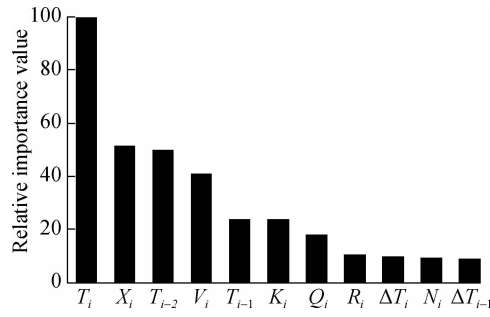

Fig. 3 The relative importance value

In the GBDT model, the partial dependency value (PDV) of the prediction model and each variable on the prediction results are shown in Fig. 4. From Fig. 4, each variable has a highly nonlinear relationship with the predicted travel time. Taking T_i as an example, when $T_i < 60$ s, the PDV for the predicted travel time is the smallest fixed value, and the change of T_i has little effect on the predicted travel time. When $60\text{ s} < T_i < 65$ s, the PDV for the predicted travel time increases sharply. However, when $65\text{ s} < T_i < 180$ s, the PDV for the predicted travel time increases slowly. When $T_i > 180$ s, the PDV for the predicted travel time becomes a fixed value again. The analysis results of the GBDT model show that the influences of the variable T_i and the predicted travel time have a typical nonlinear relationship. Only when $60\text{ s} < T_i < 180$ s, the change of T_i has a greater impact on the predicted travel time.

3.4 Accuracy of the model

Fig. 5 is a comparison between the travel time of the 5th day in the test data sets and the travel time obtained with different models. As indicated in Fig. 5, the GBDT model can accurately predict the change of travel time.

4 Conclusions

1) The comparison of model prediction results shows that the error of the GBDT model is smaller than those of the BP neural network model and the SVM model.

2) In the GBDT model, travel time in current period T_i has the highest importance value. The travel time of the current period has the greatest influence on the travel time of the next period. As expected, the immediate previous traffic state will influence the traffic in the near future. The traffic state parameter in current period X_i has a greater influence on the predicted travel time, which is similar to the driving characteristics on the road. The number of vehicles in current period N_i and the time difference ΔT_{i-1} have a small influence on the predicted travel time, indicating that the number of vehicles and the time difference cannot directly affect the travel time.

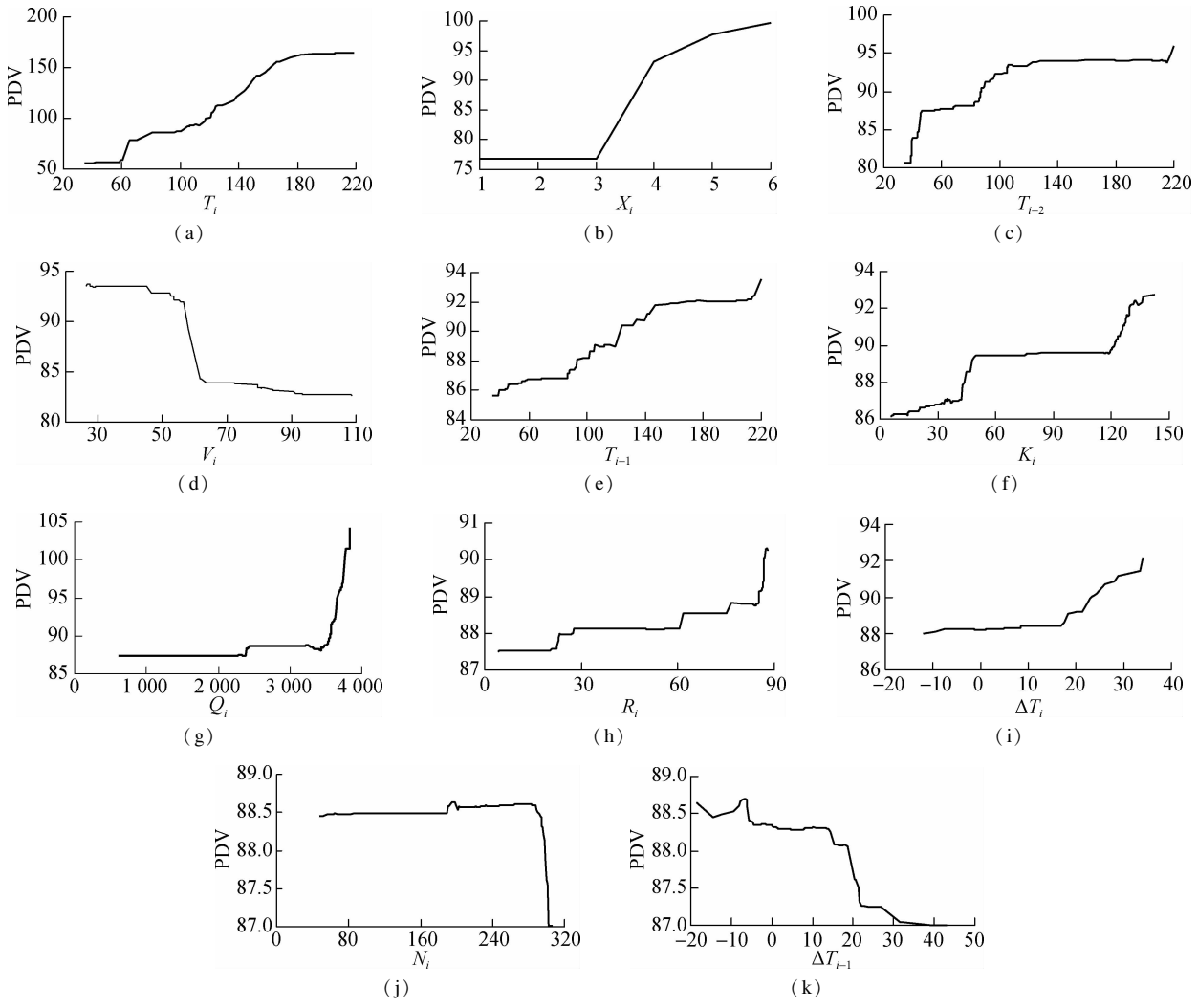

Fig. 4 The partial dependency value of each variable on the prediction result. (a) T_i; (b) X_i; (c) T_{i-2}; (d) V_i; (e) T_{i-1}; (f) K_i; (g) Q_i; (h) R_i; (i) ΔT_i; (j) N_i; (k) ΔT_{i-1}

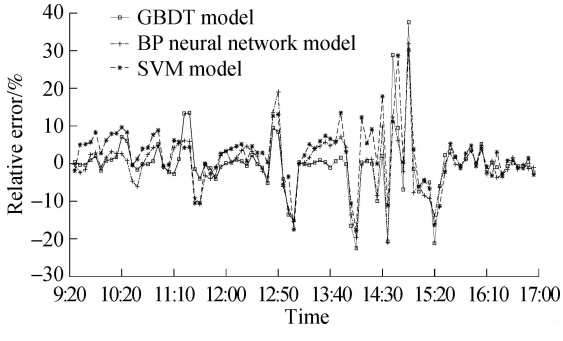

Fig. 5 Error comparison of three models

3) The partial dependency value of the variable on the prediction results indicates that the GBDT model can accurately capture the nonlinear relationship between the variables and the predicted travel time.

4) The GBDT model will be applied to other research sections for verification in the subsequent study. However, data obtained by VISSIM limits diversity. In future research, the variables of weather, characters of drivers, and other variables (holidays, working days, non-working days, morning peak, evening peak and so on) affecting travel time will be considered in the GBDT model.

References

[1] Lee W H, Tseng S S, Tsai S H. A knowledge based real-time travel time prediction system for urban network [J]. *Expert Systems with Applications*, 2009, **36**(3): 4239 - 4247. DOI:10.1016/j.eswa.2008.03.018.

[2] Miranda D M, Conceição S V. The vehicle routing problem with hard time windows and stochastic travel and service time [J]. *Expert Systems with Applications*, 2016, **64**: 104 - 116. DOI:10.1016/j.eswa.2016.07.022.

[3] Oh S, Byon Y J, Jang K, et al. Short-term travel-time prediction on highway: A review on model-based approach[J]. *KSCE Journal of Civil Engineering*, 2018, **22**(1): 298 - 310. DOI:10.1007/s12205-017-0535-8.

[4] Farokhi Sadabadi K, Hamedi M, Haghani A. Evaluating moving average techniques in short-term travel time prediction using an AVI data set[C]//*Transportation Research Board 89th Annual Meeting*. Washington, DC, USA, 2010.

[5] Julio N, Giesen R, Lizana P. Real-time prediction of bus

travel speeds using traffic shockwaves and machine learning algorithms[J]. *Research in Transportation Economics*, 2016, **59**: 250 - 257. DOI:10.1016/j.retrec.2016.07.019.

[6] Liu W M, Li S S. Freeway travel time prediction simulation research based on big data[J]. *Computer Simulation*, 2017, **34**(3): 395 - 399. (in Chinese)

[7] Cai H R, He L L. A combined offline travel time prediction model based on speed matrix and artificial neural network[J]. *Journal of Zhejiang Sci-Tech University* (*Natural Sciences Edition*), 2017(6): 851 - 858. (in Chinese)

[8] Yu B, Yang Z Z, Yao B Z. Bus arrival time prediction using support vector machines[J]. *Journal of Intelligent Transportation Systems*, 2006, **10**(4): 151 - 158. DOI: 10.1080/15472450600981009.

[9] Reddy K K, Anil Kumar B, Vanajakshi L. Bus travel time prediction under high variability conditions[J]. *Current Science*, 2016, **111**(4): 700. DOI:10.18520/cs/v111/i4/700-711.

[10] Sun X Y, Zhang H, Tian F L, et al. The use of a machine learning method to predict the real-time link travel time of open-pit trucks[J]. *Mathematical Problems in Engineering*, 2018, **2018**: 1 - 14. DOI:10.1155/2018/4368045.

[11] Kumar B A, Vanajakshi L, Subramanian S C. Bus travel time prediction using a time-space discretization approach[J]. *Transportation Research Part C: Emerging Technologies*, 2017, **79**: 308 - 332. DOI:10.1016/j.trc.2017.04.002.

[12] Zhao J D, Guo Y J, Duan X H. Dynamic path planning of emergency vehicles based on travel time prediction[J]. *Journal of Advanced Transportation*, 2017, **2017**: 1 - 14. DOI:10.1155/2017/9184891.

[13] Gal A, Mandelbaum A, Schnitzler F, et al. Traveling time prediction in scheduled transportation with journey segments[J]. *Information Systems*, 2017, **64**: 266 - 280. DOI:10.1016/j.is.2015.12.001.

[14] Zhao J D, Gao Y, Tang J J, et al. Highway travel time prediction using sparse tensor completion tactics and K-nearest neighbor pattern matching method[J]. *Journal of Advanced Transportation*, 2018, **2018**: 1 - 16. DOI:10.1155/2018/5721058.

[15] Ahmed M M, Abdel-Aty M. Application of stochastic gradient boosting technique to enhance reliability of real-time risk assessment[J]. *Transportation Research Record: Journal of the Transportation Research Board*, 2013, **2386**: 26 - 34. DOI:10.3141/2386-04.

[16] Friedman J H. Stochastic gradient boosting[J]. *Computational Statistics & Data Analysis*, 2002, **38**(4): 367 - 378. DOI:10.1016/s0167-9473(01)00065-2.

[17] Friedman J H, Meulman J J. Multiple additive regression trees with application in epidemiology[J]. *Statistics in Medicine*, 2003, **22**(9): 1365 - 1381. DOI:10.1002/sim.1501.

[18] National Research Council. *HCM*2010: *Highway capacity manual*[M]. 5th ed. Washington, DC, USA: Transportation Research Board, 2010.

[19] Zhang Y R, Haghani A. A gradient boosting method to improve travel time prediction[J]. *Transportation Research Part C: Emerging Technologies*, 2015, **58**: 308 - 324. DOI:10.1016/j.trc.2015.02.019.

基于梯度提升决策树的高速公路行程时间预测模型

程 娟 陈先华

(东南大学交通学院, 南京 211189)

摘要:为研究高速公路行程时间预测方法,基于梯度提升决策树(GBDT)建立了行程时间预测模型.提出的模型中选用 11 个变量(当前时段行程时间 T_i、当前时段流量 Q_i、当前时段速度 V_i、当前时段密度 K_i、当前时段车辆数 N_i、当前时段占有率 R_i、当前时段交通状态参数 X_i、前一个时段行程时间 T_{i-1}等)预测向前 10 min 的行程时间.利用 VISSIM 仿真得到的数据对模型进行训练和测试.结果表明,GBDT 模型的预测误差小于 BP 神经网络模型和支持向量机模型;GBDT 模型中当前时段行程时间 T_i 在所有变量中最重要.GBDT 模型能够得到更准确的预测结果,能深入挖掘变量与预测行程时间之间隐藏的非线性关系.

关键词:梯度提升决策树;行程时间预测;高速公路;交通状态参数

中图分类号:U491.2